An algorithm to improve speech recognition in noise for hearing-impaired listeners

Eric W. Healy[a] and Sarah E. Yoho
Department of Speech and Hearing Science, and Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, Ohio 43210

Yuxuan Wang and DeLiang Wang
Department of Computer Science and Engineering, and Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, Ohio 43210

(Received 20 February 2013; revised 22 August 2013; accepted 26 August 2013)

Despite considerable effort, monaural (single-microphone) algorithms capable of increasing the intelligibility of speech in noise have remained elusive. Successful development of such an algorithm is especially important for hearing-impaired (HI) listeners, given their particular difficulty in noisy backgrounds. In the current study, an algorithm based on binary masking was developed to separate speech from noise. Unlike the ideal binary mask, which requires prior knowledge of the premixed signals, the masks used to segregate speech from noise in the current study were estimated by training the algorithm on speech not used during testing. Sentences were mixed with speech-shaped noise and with babble at various signal-to-noise ratios (SNRs). Testing using normal-hearing and HI listeners indicated that intelligibility increased following processing in all conditions. These increases were larger for HI listeners, for the modulated background, and for the least-favorable SNRs. They were also often substantial, allowing several HI listeners to improve intelligibility from scores near zero to values above 70%.
 [http://dx.doi.org/10.1121/1.4820893]

PACS number(s): 43.71.Ky, 43.71.Es, 43.66.Ts, 43.72.Dv [PBN] Pages: 3029–3038

I. INTRODUCTION

A primary complaint of hearing-impaired (HI) listeners is poor speech recognition in background noise. This issue can be quite debilitating and persists despite considerable efforts to improve hearing technology. The primary limitation resulting from sensorineural hearing impairment of cochlear origin involves elevated audiometric thresholds and resulting limited audibility. Because intense sounds are often perceived at normal loudness, these listeners often have reduced dynamic range and display a steep growth of loudness as signal intensity is increased (for a review, see Bacon *et al.*, 2004). But reduced audibility forms only a portion of HI listeners' collection of limitations.

Hearing loss of 40 dB hearing level (HL) or greater is often accompanied by broad auditory tuning (for a review, see Moore, 2007). The resulting reductions in frequency resolution and spectral smearing can impair speech perception in noise (e.g., Baer and Moore, 1993; ter Keurs *et al.*, 1992). Further, HI listeners often lack the ability displayed by normal-hearing (NH) listeners to "listen in the dips" of a fluctuating masker. As a result, masking release is often reduced in HI listeners (e.g., Wilson and Carhart, 1969; Festen and Plomp, 1990; Bacon *et al.*, 1998; Bernstein and Grant, 2009), and it can be eliminated when broad tuning is simulated (ter Keurs *et al.*, 1993). It has been suggested that, in addition to a smearing of acoustic speech cues, broad tuning reduces speech recognition in complex noises by limiting opportunities to isolate spectro-temporal regions containing relatively undisturbed representations of the speech (e.g., Apoux and Healy, 2009, 2010).

Hearing-impaired listeners also display other deficits. Performance on tasks of temporal resolution is often poorer than normal. Although cochlear hearing loss may not impair temporal resolution per se, the effective resolution displayed by these listeners is often reduced due to limited audible bandwidth (e.g., Bacon and Gleitman, 1992; Moore *et al.*, 1992) or reduced sensation level (e.g., Fitzgibbons and Wightman, 1982; Glasberg *et al.*, 1987; Nelson and Thomas, 1997). It has also been suggested that an across-frequency deficit exists, in which HI listeners have particular difficulty integrating speech patterns at different spectral loci—a task presumably required to reassemble the auditory representation of a signal once decomposed by the auditory periphery (e.g., Turner *et al.*, 1999; Healy and Bacon, 2002; Souza and Boike, 2006; Grant *et al.*, 2007; Healy and Carson, 2010). Finally, and most recently, it has been suggested that HI listeners benefit less than normal from the temporal fine structure of speech (e.g., Lorenzi *et al.*, 2006).

Modern hearing aids do quite well amplifying sounds in a fashion that is appropriate for individual ears, and as a result, speech recognition in quiet can be reasonably good for many HI listeners. However, modern devices are limited in their ability to address limitations other than audibility. These limitations combine to make speech perception in noise difficult for HI listeners and the remediation of this issue equally difficult.

[a]Author to whom correspondence should be addressed. Electronic mail: healy.66@osu.edu

One technique incorporated into modern hearing technology to improve speech perception in noise involves microphone arrays. Spatial filtering, or beamforming, boosts the sound originating from a specific direction and attenuates sounds originating from other directions. The simplest implementation (delay-and-sum) assumes that the signal of interest is at zero azimuth and noise originates from elsewhere. Adaptive beamforming attempts to cancel a noise source picked up by a microphone by subtracting it from a main microphone that picks up both the target speech and the noise.

While microphone arrays can produce substantial improvements in speech-in-noise intelligibility, they are not free from limitations. First, improvement in signal-to-noise ratio (SNR) requires that target speech and interfering sounds come from different spatial locations, a rule that is often violated in natural environments. Another limitation is that of *configuration stationarity*: All spatial filtering methods operate on the premise of a fixed-source configuration (Wang, 2005). As a result, they have difficulty when sources change location or when the sound of interest switches from one source to another. These limitations together restrict situations in which hearing aids employing spatial techniques can provide benefit.

A longstanding goal in signal processing is the development of processing algorithms capable of monaural (i.e., speech and noise picked up by the same single microphone) segregation of speech from noise. Many such enhancement techniques have been proposed to perform segregation using monaural input (see Loizou, 2007). They are generally based on statistical analysis of speech and noise, followed by estimation of clean speech from noisy speech. Classic approaches include spectral subtraction, Wiener filtering, and mean-square error estimation. Spectral subtraction subtracts the power spectral density of the estimated interference from that of the mixture. The Wiener filter estimates clean speech from the ratios of speech spectrum and mixture spectrum. Mean-square error estimation models speech and noise spectra as statistically independent Gaussian random variables and estimates clean speech accordingly.

These speech-enhancement techniques can result in significant SNR increases and improved performance of automatic speech-recognition systems. However, increases in speech intelligibility for human listeners have remained elusive for decades (e.g., Levitt, 2001; Hu and Loizou, 2007). One possible reason for this lack of intelligibility improvement involves the fact that speech following separation from noise is often distorted. These processing artifacts include the well-known "musical noise" caused by spectral subtraction. Another possibility involves the removal of low-intensity speech sounds, e.g., unvoiced consonants, which are important for intelligibility. This persistent lack of success in obtaining intelligibility improvement has led some to question whether one-microphone solutions are ever possible. For example, Levitt stated that, "Our understanding of this problem is so limited that we have not only been unsuccessful in finding a solution, but we do not even know whether it is possible to improve the intelligibility of speech in noise by any significant amount." (Levitt, 1997, p. xvii).

In computational auditory scene analysis (CASA), it has been suggested that a target goal for the segregation of speech from noise is provided by the ideal binary time-frequency mask (Hu and Wang, 2001; Wang, 2005). The idea underlying the ideal binary mask (IBM) is to retain the time-frequency (T-F) regions of a mixture in which the target speech is relatively strong, and to discard the remaining regions. Specifically, the IBM is a binary matrix having a value of 1 for each T-F unit in which the SNR exceeds a threshold [or local criterion (LC)], and 0 otherwise. It is "ideal" because the mask is defined in terms of the premixed target and interference, i.e., prior knowledge of the target speech and noise is required. The term also reflects the fact that the IBM provides the optimal SNR gain of all binary T-F masks under certain conditions (Li and Wang, 2009).

A series of experiments have shown that the IBM can substantially improve intelligibility. Brungart *et al.* (2006) found that NH listeners could achieve near-perfect intelligibility in one- to three-talker interference. Anzalone *et al.* (2006) observed substantial speech reception threshold (SRT) improvements for both NH and HI listeners. Li and Loizou (2008) found NH-intelligibility results broadly consistent with those of Brungart *et al.* (2006). Wang *et al.* (2009) observed considerable SRT improvements for both NH and HI listeners in speech-shaped noise (SSN) and in cafeteria noise. The larger improvements in the latter complex background suggest that ideal masking is more effective for modulated than for stationary noise. Further, Wang *et al.* (2009) found that ideal masking was capable of raising the intelligibility of HI listeners to levels comparable to that of NH listeners. Finally, Cao *et al.* (2011) showed that adding background noise to fill in 0-valued T-F units in the IBM can further improve intelligibility.

Although binary masking is clearly capable of producing large intelligibility gains, to be useful, a separation algorithm must be able to *estimate* the IBM directly from a noisy mixture, i.e., without prior knowledge of the individual target and noise signals. To our knowledge, the only such demonstration of intelligibility improvement from a monaural algorithm is provided by Kim *et al.* (2009). The authors proposed an algorithm that uses a Gaussian mixture model (GMM) classifier to decide whether each T-F unit is dominated by speech or by noise. Unlike the IBM, the binary mask employed by Kim *et al.* (2009) was estimated by training the GMM classifier. Sentences from the IEEE database (IEEE, 1969) were separated in this way from one of three noises (20-talker babble, factory noise, or SSN) at −5 or 0 dB SNR. Results from NH listeners indicated improvements in intelligibility in all conditions and substantial improvements when unprocessed scores were low. An adapted version of this algorithm was later shown to improve speech intelligibility of cochlear implant (CI) users (Hu and Loizou, 2010). Recently, Fink *et al.* (2012) tested a different binary-masking algorithm tailored specifically for white noise, and found a clear intelligibility benefit for CI users, but not for NH listeners or hearing-aid users.

The results of Kim *et al.* (2009) are impressive, but they are restricted to NH listeners or CI users. For an algorithm to be useful to the largest population of hearing-impaired

listeners—those who need hearing aids—it must be capable of improving intelligibility for such listeners. Further, from the algorithmic standpoint, GMM classifiers tend to overfit the training set. As a result, the Kim *et al.* algorithm likely has difficulty handling even small noise variations between training and test conditions (Han and Wang, 2012; Wang and Wang, 2013). The goal of the current study is to evaluate a new binary-masking algorithm designed to improve intelligibility for both NH and HI listeners. Sentences from the Hearing in Noise Test (HINT) (Nilsson *et al.*, 1994) were presented in steady noise and in babble at various SNRs, prior to and after processing, to both types of listeners.

II. ALGORITHM DESCRIPTION

Due to the success of the IBM in improving speech intelligibility for both NH and HI listeners, we approach the segregation of speech from noise through IBM estimation. In other words, speech segregation is treated as a binary-classification problem in which each T-F unit needs to be labeled as speech-dominant (1) or noise-dominant (0). In the current study, the IBM is estimated by training using sentences not used for testing. Figure 1 shows a schematic diagram of the current system. Noisy signals were first passed through a 64-channel gammatone filterbank with center frequencies ranging from 50 to 8000 Hz. The output from each filter channel was divided into 20-ms frames with 10-ms overlap. This formed a T-F representation known as a cochleagram (Wang and Brown, 2006), from which acoustic features were extracted. During the training stage, the IBM provided binary labels reflecting speech or noise dominance in each T-F unit. Using the standard training procedure of backpropagation, the estimated IBM was obtained by minimizing the difference between it and the IBM. In supervised learning, both feature extraction and classifier training are important, and they are discussed separately below.

It was also important to set a proper value for the LC, which again is the SNR criterion used to label a particular T-F unit as speech-dominated or noise-dominated. In the current study, the following values for LC were used: −6 dB for input SNRs of 0 and −2 dB, −10 dB for SNR of −5 dB, and −12 dB for SNR of −8 dB.

A. Feature extraction

Since a binary decision needed to be made for each T-F unit, acoustic features were extracted from each T-F unit. Wang *et al.* (2013) conducted a systematic evaluation of different unit-level features and identified a set of complementary features. In the current study, this complementary feature set was employed, which consisted of (1) the amplitude modulation spectrogram (AMS), (2) relative spectral transform and perceptual linear prediction (RASTA-PLP), and (3) mel-frequency cepstral coefficients (MFCCs). Although each of these three feature types can be used to discriminate speech from noise to some degree, the use of all three adds discriminative power (Wang *et al.*, 2013).

The procedure of Kim *et al.* (2009) was employed to extract the 15-dimensional (15-D) AMS feature, which is composed of 15 modulation-frequency bins. Briefly, the envelope within each spectral-frequency channel was extracted using full-wave rectification followed by decimation. The modulation spectrum was then obtained by passing the Hanning-windowed decimated envelope to a 256-point FFT. Finally, the 256 FFT modulation magnitudes were reduced to 15 values using 15 triangular averaging windows.

The extraction of RASTA-PLP and MFCC features followed common practice. To extract the 13-D RASTA-PLP feature, the power spectrum was first warped to the Bark scale, which was then log compressed. This auditory spectrum was then filtered by the RASTA filter and expanded again by an exponential function. Finally, the PLP analysis was performed on this filtered spectrum. To extract the 31-D MFCC feature, the signal was first preemphasized, then a 512-point FFT with a 20-ms Hamming window was used to obtain its power spectrum. The power spectra were then warped to the mel scale followed by the standard log operation and discrete cosine transform.

Incorporating Δ features has been found to provide significant improvements in classification. These features are simply difference values between neighboring T-F units, which capture the temporal variation of a feature. To balance performance with computational overhead, first- and second-order Δ features were used only for RASTA-PLP. In total, these features together resulted in an 85-D feature vector for each T-F unit.

B. Classifier training

Previous classification-based segregation systems have employed GMM (Kim *et al.*, 2009) or support vector machines (SVM, Han and Wang, 2012). Wang and Wang (2013) showed that deep neural networks (DNNs) outperform both of these. DNNs were therefore employed as classifiers in the current study. Because they operate within each frequency channel, they are referred to as subband DNN classifiers in Fig. 1.

DNNs generally refer to neural networks having more than one hidden layer. They can be viewed as hierarchical

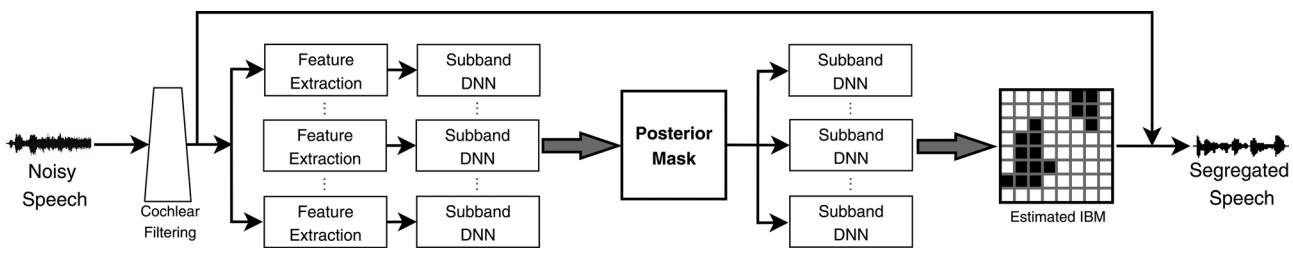

FIG. 1. Schematic diagram of the current speech-segregation system. DNN = deep neural network, IBM = ideal binary mask.

feature detectors that increasingly untangle factors of variation as the number of hidden layers increases. However, training with more than one hidden layer was previously considered difficult due to problems such as vanishing gradients. The resulting models either had high training errors or they overfit a particular training set. To address these problems, Hinton *et al*. (2006) proposed to first pre-train the network, and then use supervised training to fine tune the network. Specifically, they proposed to use a stack of restricted Boltzmann machines (RBMs), trained in an unsupervised layerwise fashion for pre-training. An RBM is a two-layer network trained to model its input data. After RBM pre-training, the resulting weights (the tunable parameters) were used as initial values for standard backpropagation training. The main advantage of DNNs lies in unsupervised RBM pre-training, which aims to represent the input data in increasingly more abstract ways to effectively encode stable (robust) features in the data and enable improved discriminative training via backpropagation.

In the current study, the 85-D acoustic features described above were used as raw inputs to DNNs. We found that using two hidden layers with RBM pre-training significantly improved classification performance, with more layers providing diminished performance gain. A Gaussian–Bernoulli RBM was used for the first hidden layer to deal with real-valued inputs, and a Bernoulli–Bernoulli RBM was used for the second hidden layer. Both hidden layers had 200 units, each of which used a logistic sigmoid transfer function. One iteration of contrastive divergence (Hinton, 2002) was used to approximate the gradient in RBM, and learning rates of 0.01 and 0.1 were used for training the Gaussian–Bernoulli and Bernoulli–Bernoulli RBM, respectively.

Following this pre-training, fine tuning of the DNNs took place using the backpropagation procedure, in which the cross entropy objective function was used to measure the error with respect to the IBM. In both RBM pre-training and backpropagation learning, mini-batch gradient descent with a batch size of 512 was used. The interested reader is referred to Hinton *et al*. (2006) and Wang and Wang (2013) for more technical discussions of the learning algorithms.

C. Incorporating contextual information

As described above, the labeling of each T-F unit was based on its acoustic features. However, speech typically exhibits highly structured spectro-temporal patterns that result from the human speech-production mechanism. Because that mechanism possesses constraints and mechanical inertias, and because languages introduce additional constraints due to various rules, these patterns tend to be structured. Therefore, taking into consideration acoustic features from neighboring T-F units is expected to benefit decision making in the current T-F unit. However, direct concatenation of raw features results in very high-dimensional feature vectors, which may render training impractical. To alleviate this concern, a first DNN was trained as described above, which output a posterior probability of target-dominance for each T-F unit (posterior mask in Fig. 1). Then, a second DNN was trained, in which a window of posterior probabilities was concatenated as the representation for the center T-F unit. A window spanning both five time frames and 17 (of the 64) frequency channels was used. Such a representation was both discriminative and parsimonious, and we have found significant improvements in classification by incorporating this contextual information.

Figure 2 illustrates the segregation of a HINT utterance from SSN at −5 dB SNR. Panels (a) and (b) show the cochleagrams of the clean speech and speech-plus-noise, respectively. The IBM, estimated IBM and cochleagram of the segregated speech are shown in panels (c), (d), and (e), respectively.

III. HUMAN SUBJECTS TESTING

A. Method

1. Subjects

Twelve listeners diagnosed with a bilateral sensorineural hearing loss of cochlear origin participated. These individuals were selected to represent typical HI listeners seen at The Ohio State University Speech-Language-Hearing Clinic. All were bilateral hearing aid users recruited from this clinic. The prior diagnoses were confirmed on day of test using pure-tone audiometry (ANSI, 2004) and tympanometry (ANSI, 1987). Audiograms generated on day of test are displayed in Fig. 3. Although the hearing losses may be characterized, on average, as sloping and moderate, they ranged from flat to sloping and from mild to severe. Also displayed in Fig. 3 are pure-tone averages (PTAs) based on 0.5, 1, and 2 kHz and pooled across ears (range = 33–54 dB HL, mean = 42.8), genders (seven females), and subject ages (range = 32–72 yr, mean = 61.8).

In addition to this group, 12 NH listeners were recruited from undergraduate courses at The Ohio State University. All had pure-tone audiometric thresholds of 20 dB HL or better at octave frequencies from 250 to 8000 Hz (ANSI, 2004). They were aged 19 to 28 yr (mean = 21.1), and all were female. All subjects received a monetary incentive or course credit for participating.

2. Stimuli and procedure

The original male-talker, 20 161-Hz, 16-bit digital recordings of the HINT sentences described in Nilsson *et al*. (1994) were employed. Prior to processing, the signals were downsampled to 16 kHz and each sentence was scaled to equate total RMS energy. The algorithm was trained using 100 sentences (sentences 1–70 and 251–280) and subjects were tested using 160 different sentences (sentences 71–230).

The two noise backgrounds included SSN and multi-talker babble. The SSN was that from the commercial version of the HINT and was 10 s in duration. The multi-talker babble was created by mixing at equal amplitudes sentences from the TIMIT database (Garofolo *et al*., 1993) spoken by eight different talkers (four male and four female talkers, two sentences each). The sentences were mixed with SSN at −2, −5, or −8 dB SNR and with babble at 0, −2, or −5 dB SNR, where SNR was calculated over the duration of a HINT sentence. Each training and test sentence was mixed

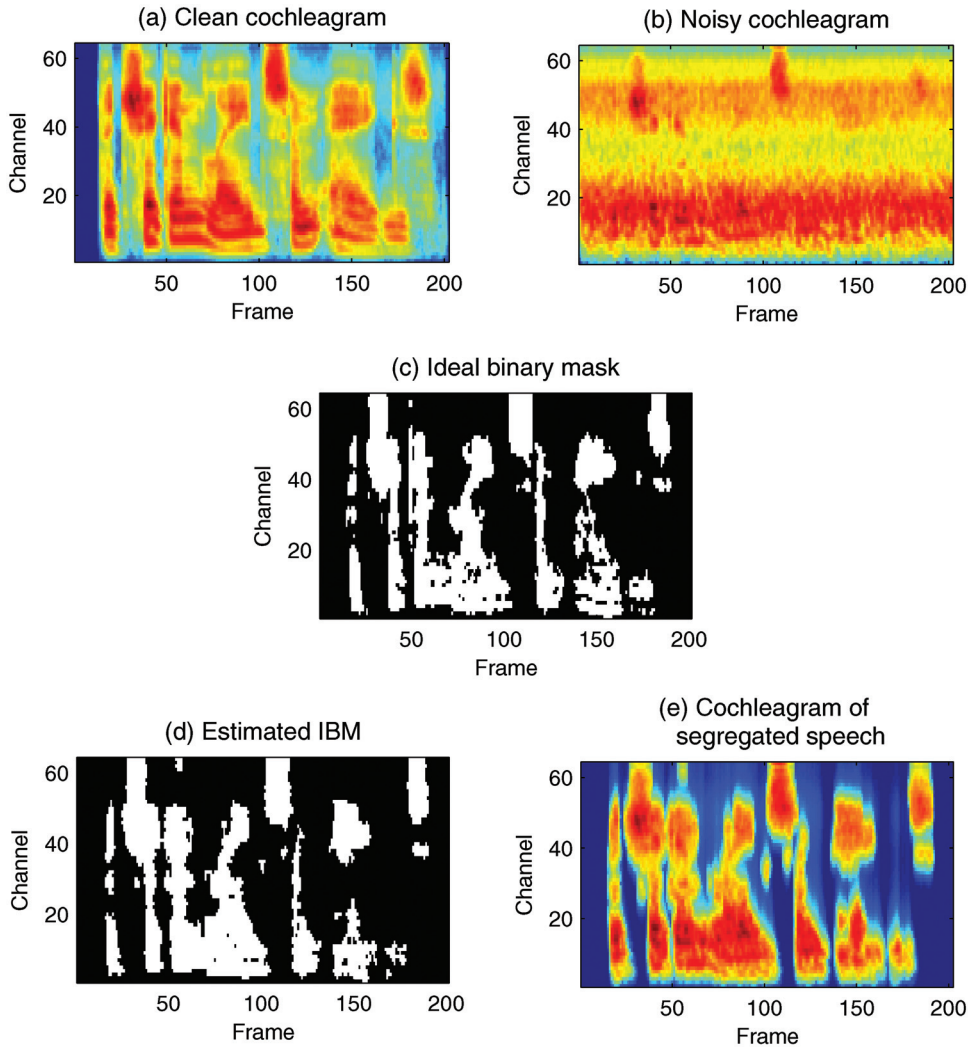

FIG. 2. Segregation of a HINT utterance from speech-shaped noise at −5 dB SNR. (a) Cochleagram of the utterance in quiet. (b) Cochleagram of the speech-plus-noise. (c) IBM. (d) Estimated IBM. (e) Cochleagram of the utterance following noise removal using the estimated IBM.

with a sample of SSN or babble having a randomly determined start point within the looped noise. The noise began approximately 140 ms prior to the beginning of each sentence and ended approximately 140 ms after the end of each sentence.

Testing began with audiometric evaluation followed by a brief familiarization. This familiarization consisted of five HINT sentences in each of the following conditions in the following order: (1) unprocessed in quiet, (2) unprocessed in SSN, (3) processed in SSN, (4) unprocessed in babble, and (5) processed in babble. Unprocessed refers to the original speech or speech-noise mixture and processed refers to this same mixture following processing by the current algorithm. Five sentences used for training were presented during stage (1), and sentences not used for training or test were used for stages (2–5). SNR was set to 0 dB during familiarization. Following this familiarization, each listener heard 20 HINT sentences in each of eight conditions (2 processed/unprocessed × 2 SSN/babble × 2 SNRs). Each subject group (NH or HI) heard two of the three SNRs for each noise type. The sentence list-to-condition correspondence was pseudorandomized for each subject. The presentation order of conditions was also pseudorandomized for each subject, with the restriction that unprocessed/processed conditions appeared successively in random order (i.e., either unprocessed first or processed first) for a given SNR and noise type.

The signals were presented diotically over Sennheiser HD 280 headphones (Wedemark, Germany) using a personal computer equipped with Echo Digital Audio (Santa Barbara, CA) Gina 3G digital-to-analog converters. Presentation levels were set using a Larson Davis (Depew, NY) sound level meter and flat-plate coupler (models 824 and AEC 101). The average RMS level of continuous speech or speech-plus-noise was set to 65 dBA for the NH listeners. The presentation level for the HI listeners (tested with hearing aids removed) was set initially to 85 dBA. Following the first five unprocessed familiarization sentences, HI subjects were asked if the presentation level was adequate and comfortable and, "if they would turn it up or down if they could." All but one subject reported that the initial presentation level was adequate and comfortable. For the one subject who desired an increase in level (HI3), the presentation level was increased to 90 dBA, which was judged to be adequate and

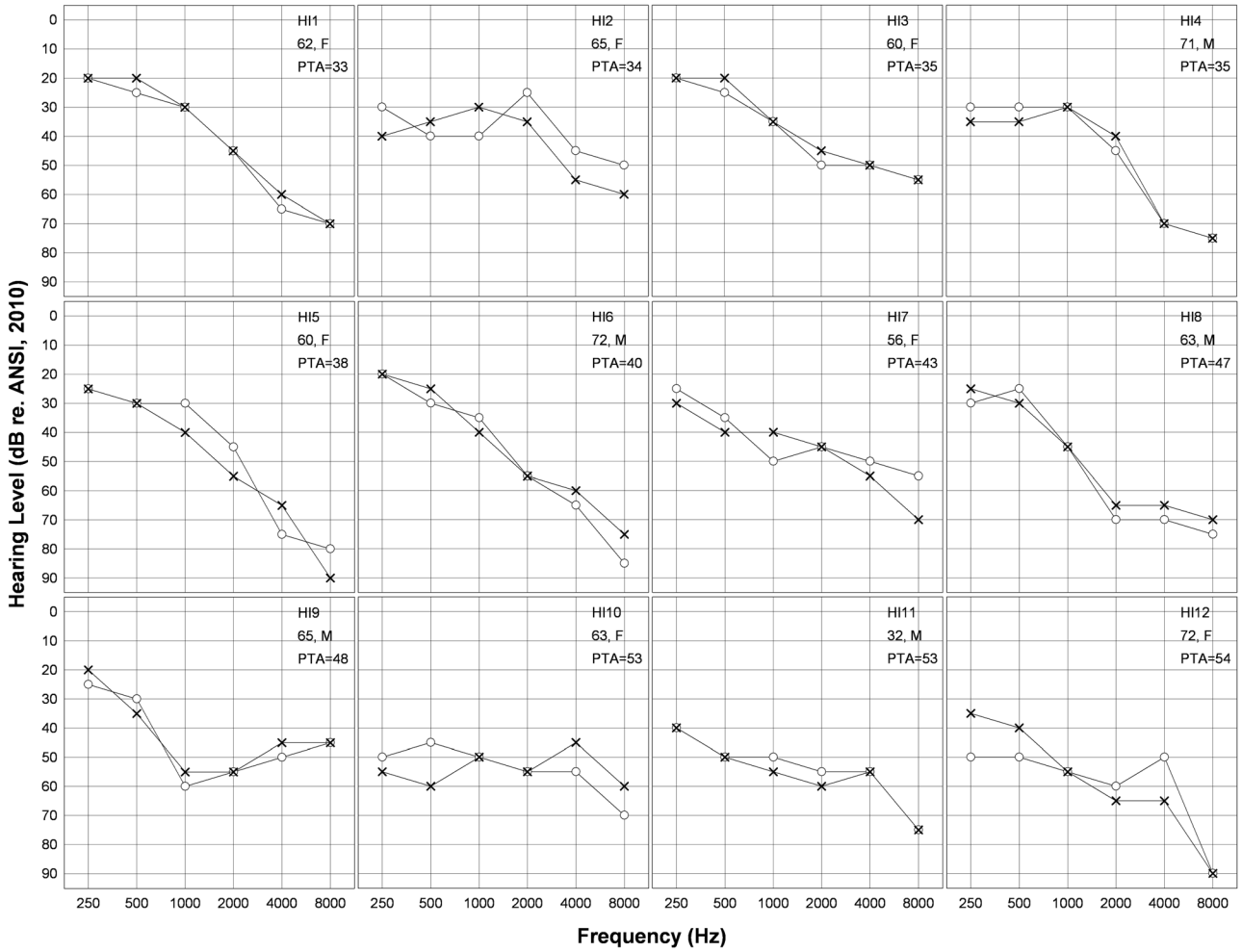

FIG. 3. Pure-tone air-conduction audiometric thresholds for the listeners with sensorineural hearing impairment. Thresholds in right ears are represented by circles and those in left ears are represented by ×'s. Also displayed are listener ages in years and genders, as well as PTAs (in dB HL based on thresholds at 0.5, 1, and 2 kHz and pooled across ears). Listeners are numbered and arranged according to increasing PTA.

comfortable when the initial five familiarization sentences were repeated.

Subjects were seated with the experimenter in a double-walled audiometric booth. They repeated back as much of each sentence as they could, and the experimenter recorded the number of words correctly reported. Scoring was based on percentage of component words correctly recalled.

B. Results and discussion

The mean intelligibility for each subject in each condition is displayed in Fig. 4. The upper panels display data for sentences in SSN and the lower panels display data for sentences in babble. For each listener, the unprocessed score is represented by a circle and the processed score is represented by a triangle. The benefit of processing is therefore represented by the height of the bar connecting the two symbols. It is clear from Fig. 4 that both NH and HI subjects demonstrated improvements in intelligibility following processing. Individual HI listeners demonstrated the largest gains. In SSN at −5 dB, the HI subject who displayed single-digit intelligibility scores in the unprocessed condition increased to 77% when processed. In babble at −2 dB, three HI subjects displayed single-digit scores in the unprocessed condition and recognition of 71–85 % when processed. Two more HI subjects displayed unprocessed scores below 15% and recognition of 81–86 % when processed.

Figure 5 displays group mean performance for each noise type, SNR, and listener group. Intelligibility for the NH listeners increased from 36.7 to 80.1 % in the least-favorable SSN and from 42.3 to 77.8 % in the least-favorable babble. Intelligibility for the HI listeners increased from 35.9 to 81.7 % in the least-favorable SSN and from 28.6 to 83.6 % in the least-favorable babble. A series of planned comparisons (uncorrected paired t tests) confirmed the reliability of the processing benefit in each condition shown in Fig. 5 [$t(11) \geq 4.9$, $p < 0.001$].[1]

The benefit displayed by the HI listeners was generally larger than that displayed by the NH listeners. This difference between listeners is larger in the babble background and is most apparent at the common SNR of −2 dB (bottom center panel of Fig. 5). The benefit advantage for the HI listeners remains when benefit is compared across SNRs that produced comparable unprocessed scores: In babble, unprocessed scores averaging 42 and 40 % (at −5 dB for NH and 0 dB for HI, respectively) rose to 78 and 90 % when processed, resulting in mean benefits of 35% for NH versus 50% for HI [$t(22) = 2.1$, $p < 0.05$]. The HI benefit advantage is also apparent in SSN at

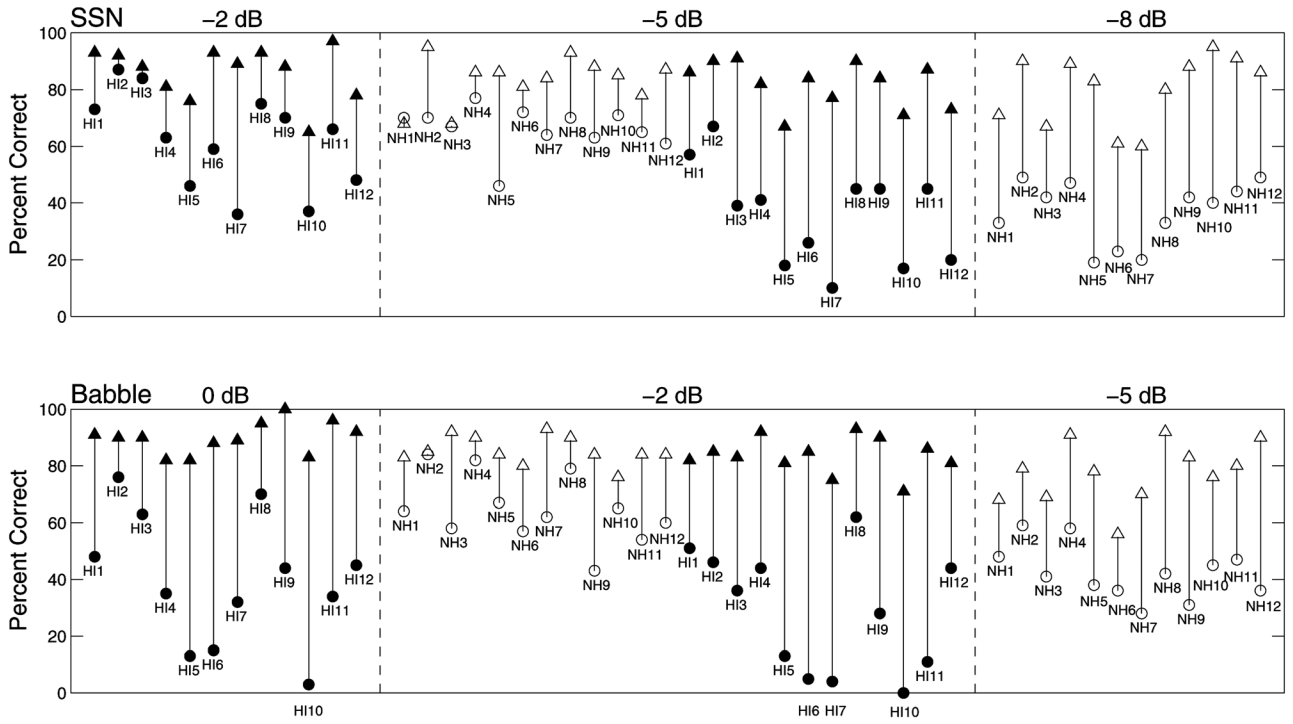

FIG. 4. Mean HINT sentence component-word recognition scores for each listener. Normal-hearing listeners are represented by open symbols and hearing-impaired listeners are represented by filled symbols. Unprocessed conditions are represented by circles, and algorithm-processed conditions are represented by triangles. The upper panels display recognition in speech-shaped noise at three SNRs indicated, and the lower panels display recognition in multi-talker babble at three SNRs. The hearing-impaired listeners are numbered and plotted in order of increasing pure-tone average.

the common SNR of −5 dB (top center panel of Fig. 5). It remains, to a more modest degree than for babble, when conditions that produced comparable unprocessed scores are compared. Comparison between the two right-most pairs of columns in the upper panels of Fig. 5 (SNR's of −5 versus −8 dB), or comparison between the two left-most pairs of columns (SNR's of −2 versus −5 dB), indicates that benefit was slightly larger for HI than for NH listeners.

Another comparison of interest involves performance of the NH listeners prior to processing versus that of the HI listeners following processing, in conditions of common SNR (Fig. 5, center panels). It was found that the HI listeners hearing processed stimuli significantly outperformed the NH listeners hearing unprocessed stimuli in SSN [81.7 versus 66.4 %, $t(22) = 4.8$, $p < 0.001$], and in babble [83.6 versus 64.7 %, $t(22) = 4.7$, $p < 0.001$].

IV. GENERAL DISCUSSION

Figure 4 shows that intelligibility in the processed conditions was relatively homogeneous across individual HI listeners, whereas intelligibility in the corresponding unprocessed conditions was far more heterogeneous. Thus, benefit was determined largely by performance in the unprocessed conditions. The heterogeneity in unprocessed scores is to be expected. However, the homogeneously high individual scores in the processed conditions, despite large differences in unprocessed scores, indicate that the current algorithm is capable of outputting speech that is intelligible for HI listeners who vary widely in speech-in-noise performance, at least for these speech materials.

Figure 5 shows that group-mean intelligibility scores in the unprocessed conditions were markedly reduced as SNR was reduced, as expected. However, mean intelligibility across processed conditions was quite stable. The homogeneously high mean scores across the processed conditions indicate that the current algorithm is capable of outputting speech that is intelligible for NH and HI listeners across a range of SNR values.

Relationships between various subject variables and benefit were examined in an attempt to identify HI-subject characteristics related to maximum benefit. No correlations were observed between subject age and benefit (or age and raw unprocessed or processed scores). Instead, listeners displayed considerable benefit across the range of ages tested. In contrast, relationships were observed between benefit and amount of hearing loss. The HI listeners in Fig. 4 are arranged in order of increasing PTA. Thus, the mildest hearing impairments in the center panels are juxtaposed with the NH data. As should be expected, unprocessed scores tended to be higher for listeners having lower (better) PTAs. Benefit tended to be related to PTA as a consequence of this systematic relationship between unprocessed scores and PTA. However, all listeners tended to produce high intelligibility in the processed conditions, regardless of degree of hearing loss. Thus, the current algorithm produced the maximum benefit for the listeners who needed it most—those who performed most poorly in background noise. While this trend is evident in each set of HI data in Fig. 4,

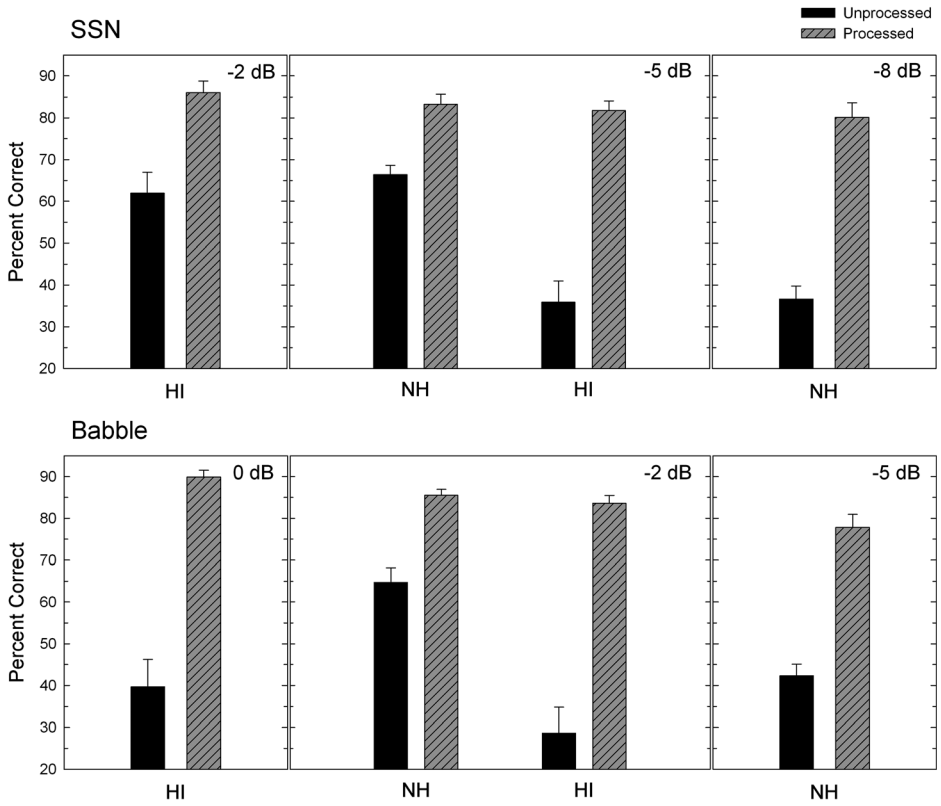

FIG. 5. Group mean component-word recognition scores and standard errors for HINT sentences presented in speech-shaped noise (upper panels) and multi-talker babble (lower panels), at the SNRs indicated, for normal-hearing and hearing-impaired listeners, both prior to and following algorithm processing.

the correlation values did not reach statistical significance, perhaps due to the limited number of samples and limited power of the tests.

The current study, along with Kim *et al*. (2009), clearly confirms the utility of binary classification for improving speech intelligibility in noise. In addition to increasing intelligibility for NH listeners, the current study demonstrates that our algorithm is capable of outputting speech information that is sufficient for an impaired auditory system, which typically has reduced dynamic range, poor frequency resolution, and effectively reduced temporal resolution, as well as other potential limitations. Since the speech material and noises used in the current study are different from those used in Kim *et al*., the amounts of improvement should not be directly compared. Perhaps a more meaningful comparison can be drawn in terms of the quality of IBM estimation, which is the goal of both their algorithm and ours. By analyzing the correlation between objective classification results and speech-intelligibility scores, Kim *et al*. suggest the HIT-FA rate for quantifying the performance of speech-segregation algorithms, where HIT is the percent of target-dominant T-F units (i.e., 1's in the IBM) correctly classified and FA (false alarm) is the percent of noise-dominant units incorrectly classified. By this metric, the current DNN-based algorithm obtains an average HIT-FA rate of 79.3% for SSN in the −5 dB SNR condition while their GMM-based algorithm yields 64.2 and 76.1% for the three-noise and one-noise training conditions, respectively (see male-speaker data in Table I of Kim *et al*., 2009). Although the eight-talker babble in the current study is expected to be a more difficult interference than the 20-talker babble in Kim *et al*., our algorithm obtains an 80.9% HIT-FA rate at −5 dB SNR while their corresponding rates are 59.4 and 72.4 % for the three-noise and one-noise conditions, respectively. This comparison shows that the current DNN-based classification produces better IBM estimation than the GMM-based classification of Kim *et al*. (see also Wang and Wang, 2013).

From the standpoint of improving the SNR of segregated speech, the optimal LC choice should be 0 dB (Li and Wang, 2009), which is different from the negative LC values used in this study. Indeed, our informal listening tests indicate that the choice of LC = 0 dB leads to significantly less intelligible speech. Part of the reason is that, with negative overall input SNRs, the resulting binary masks become sparse, having fewer 1's and losing more speech signal. This is consistent with previous intelligibility studies on binary masking suggesting that negative LC values are more appropriate for improving speech intelligibility (Brungart *et al*., 2006; Li and Loizou, 2008; Wang *et al*., 2009; Kim *et al*., 2009). This also indicates that maximizing SNR may be counterproductive if the objective is to improve human speech intelligibility in background noise (see also, Wang *et al*., 2009). Since SNR maximization is tantamount to producing an output signal as close as possible to the target speech, which is the implicit goal of speech-enhancement methods, this may be an important reason why such methods have failed to elevate speech intelligibility (see Sec. I).

The comparison between performance of NH listeners prior to processing and that of HI listeners following processing, in conditions of common SNR, is analogous to examining these listeners in similar acoustic environments, should the HI listeners have access to an algorithm like the one described here. The results of this comparison suggest the potential for the current algorithm to improve performance for HI listeners: The fact that HI listeners significantly outperformed NH listeners indicates that impaired listeners

have the potential to perform as well as, if not better than, their NH counterparts in challenging environments, given the current processing.

The fact that the current algorithm is capable of producing intelligibility by HI listeners that exceeds that achieved when their NH counterparts are presented with noisy stimuli is quite encouraging and suggests that the current algorithm may potentially be simplified in various ways (e.g., without using two future frames described in Sec. II C) to reduce processing demand, while still providing adequate levels of benefit. This may be important, given an eventual goal of implementation into hearing technology, including hearing aids and cochlear implants. We stress that this goal is long term and that the current algorithm is far from ready to implement. On the other hand, the current algorithm possesses attributes suggesting that its eventual implementation may be possible. First, the monaural nature of the algorithm provides inherent convenience in implementation relative to microphone-array techniques. Second, the classification-based framework shifts much of the workload to a training stage. During the operational (test) stage, the algorithm involves only feature extraction and binary labeling using trained classifiers, both of which could be performed efficiently. As an indication of processing time, the current algorithm takes approximately 123 ms (107 for feature extraction and 16 for DNN classification) per frequency channel to separate a 3-s noisy utterance using a single Intel 2.8 GHz Xeon processor. We should mention that no attempt was made to optimize processing speed as this was not an objective of the current study; e.g., a significant increase in speed could be achieved by replacing the current MATLAB implementation of feature extraction with a C implementation.

An inherent issue in supervised learning is generalization—typically a trained classifier is not expected to generalize well to completely new acoustic conditions. Like Kim *et al.* (2009), talker, SNR level, and noise types were matched across training and test stages in the current study, while speech utterances (sentence content) were varied across the two stages. As demonstrated by Kim *et al.*, talker mismatch is not a major issue. This is because classifiers are trained to distinguish between speech- and noise-dominant T-F units, and the acoustic characteristics of speech-dominant units are generally different from those of noise-dominant units, even when that noise is babble. We consider SNR mismatch to be of less concern than noise mismatch because SNR estimation can be performed with reasonable accuracy (Kim and Stern, 2008; Narayanan and Wang, 2012). Regarding noise mismatch, recent effort has been made to address this issue. In Han and Wang (2013), an adaptation technique based on voice-activity detection has been suggested to obtain glimpses of background noise during speech-absent frames. In Wang and Wang (2013), it was proposed that classifiers be trained on a large number of noises (and talkers) in order to cover a variety of background interferences during training. Although these techniques help to alleviate the generalization issue, their effectiveness in improving speech intelligibility in arbitrary environments remains to be tested. Clearly, generalization will be an important issue for future research.

To summarize, the current results indicate substantial increases in sentence component-word intelligibility resulting from a monaural speech-segregation algorithm. The increase is apparent for both NH and for HI listeners, and is largest in modulated backgrounds and for HI listeners. To our knowledge, this is the first monaural algorithm that provides demonstrated speech intelligibility improvements for HI listeners in background noise.

ACKNOWLEDGMENTS

This work was supported in part by grants from the National Institute on Deafness and other Communication Disorders (Grant No. R01 DC08594 to E.W.H. and Grant No. R01 DC012048 to D.L.W.) and from the Air Force Office of Scientific Research (Grant No. FA9550-12-1-0130 to D.L.W.) and an STTR subcontract from Kuzer (to D.L.W.).

[1]All planned-comparison t tests yielding $p < 0.001$ would yield significant differences following Bonferroni correction for multiple-comparisons.

ANSI (**1987**). S3.39 (R2012), *American National Standard Specifications for Instruments to Measure Aural Acoustic Impedance and Admittance (Aural Acoustic Immittance)* (Acoustical Society of America, New York).

ANSI (**2004**). S3.21 (R2009), *American National Standard Methods for Manual Pure-Tone Threshold Audiometry* (Acoustical Society of America, New York).

Anzalone, M. C., Calandruccio, L., Doherty, K. A., and Carney, L. H. (**2006**). "Determination of the potential benefit of time-frequency gain manipulation," Ear Hear. **27**, 480–492.

Apoux, F., and Healy, E. W. (**2009**). "On the number of auditory filter outputs needed to understand speech: Further evidence for auditory channel independence," Hear. Res. **255**, 99–108.

Apoux, F., and Healy, E. W. (**2010**). "Relative contribution of off- and on-frequency spectral components of background noise to the masking of unprocessed and vocoded speech," J. Acoust. Soc. Am. **128**, 2075–2084.

Bacon, S. P., Fay, R. R., and Popper, A. N. (**2004**). *Compression: From Cochlea to Cochlear Implants* (Springer, New York), pp. 136–152.

Bacon, S. P., and Gleitman, R. M. (**1992**). "Modulation detection in subjects with relatively flat hearing losses," J. Speech Hear. Res. **35**, 642–653.

Bacon, S. P., Opie, J. M., and Montoya, D. Y. (**1998**). "The effects of hearing loss and noise masking on the masking release for speech in temporally complex backgrounds," J. Speech Lang. Hear. Res. **41**, 549–563.

Baer, T., and Moore, B. C. J. (**1993**). "Effects of spectral smearing on the intelligibility of sentences in noise," J. Acoust. Soc. Am. **94**, 1229–1241.

Bernstein, J. G. W., and Grant, K. W. (**2009**). "Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **125**, 3358–3372.

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.

Cao, S., Li, L., and Wu, X. (**2011**). "Improvement of intelligibility of ideal binary-masked noisy speech by adding background noise," J. Acoust. Soc. Am. **129**, 2227–2236.

Festen, J. M., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.

Fink, N., Furst, M., and Muchnik, C. (**2012**). "Improving word recognition in noise among hearing-impaired subjects with a single-channel cochlear noise-reduction algorithm," J. Acoust. Soc. Am. **132**, 1718–1731.

Fitzgibbons, P. L., and Wightman, F. L. (**1982**). "Gap detection in normal and hearing-impaired listeners," J. Acoust. Soc. Am. **72**, 761–765.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (**1993**). "DARPA TIMIT acoustic-phonetic continuous speech corpus," technical report No. NISTIR4930, National Institute of Standards and Technology.

Glasberg, B. R., Moore, B. C. J., and Bacon, S. P. (**1987**). "Gap detection and masking in hearing-impaired and normal-hearing subjects," J. Acoust. Soc. Am. **81**, 1546–1556.
Grant, K. W., Tufts, J. B., and Greenberg, S. (**2007**). "Integration efficiency for speech perception within and across sensory modalities by normal-hearing and hearing-impaired individuals," J. Acoust. Soc. Am. **121**, 1164–1176.
Han, K., and Wang, D. (**2012**). "A classification based approach to speech segregation," J. Acoust. Soc. Am. **132**, 3475–3483.
Han, K., and Wang, D. L. (**2013**). "Towards generalizing classification based speech separation," IEEE Trans. Audio Speech Lang. Process. **21**, 166–175.
Healy, E. W., and Bacon, S. P. (**2002**). "Across-frequency comparison of temporal speech information by listeners with normal and impaired hearing," J. Speech Lang. Hear. Res. **45**, 1262–1275.
Healy, E. W., and Carson, K. A. (**2010**). "Influence of broad auditory tuning on across-frequency integration of speech patterns," J. Speech Lang. Hear. Res. **53**, 1087–1095.
Hinton, G. E. (**2002**). "Training products of experts by minimizing contrastive divergence," Neural Comput. **14**, 1771–1800.
Hinton, G. E., Osindero, S., and Teh, Y.-W. (**2006**). "A fast learning algorithm for deep belief nets," Neural Comput. **18**, 1527–1554.
Hu, G., and Wang, D. L. (**2001**). "Speech segregation based on pitch tracking and amplitude modulation" in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 79–82.
Hu, Y., and Loizou, P. C. (**2007**). "A comparative intelligibility study of single-microphone noise reduction algorithms," J. Acoust. Soc. Am. **122**, 1777–1786.
Hu, Y., and Loizou, P. C. (**2010**). "Environment-specific noise suppression for improved speech intelligibility by cochlear implant users," J. Acoust. Soc. Am. **127**, 3689–3695.
IEEE (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.
Kim, C., and Stern, R. M. (**2008**). "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Proceedings of INTERSPEECH*, pp. 2598–2601.
Kim, G., Lu, Y., Hu, Y., and Loizou, P. C. (**2009**). "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," J. Acoust. Soc. Am. **126**, 1486–1494.
Levitt, H. (**1997**). "NEW TRENDS: Digital hearing aids: Past, present, and future," *Guest Editorial in Practical Hearing Aid Selection and Fitting*, edited by H. Tobin (VA-Rehabilitation R&D Service, Washington DC), pp. xi–xxiii.
Levitt, H. (**2001**). "Noise reduction in hearing aids: A review," J. Rehab. Res. Dev. **38**, 111–121.
Li, N., and Loizou, P. C. (**2008**). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," J. Acoust. Soc. Am. **123**, 1673–1682.
Li, Y., and Wang, D. L. (**2009**). "On the optimality of ideal binary time-frequency masks," Speech Commun. **51**, 230–239.
Loizou, P. C. (**2007**). *Speech Enhancement: Theory and Practice* (CRC Press, Boca Raton, FL), Chap. 5–8.
Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. C. J. (**2006**). "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," Proc. Natl. Acad. Sci. U.S.A. **103**, 18866–18869.
Moore, B. C. J. (**2007**). *Cochlear Hearing Loss: Physiological, Psychological and Technical Issues*, 2nd ed. (Wiley, Chichester), pp. 45–91.
Moore, B. C. J., Shailer, M. J., and Schooneveldt, G. P. (**1992**). "Temporal modulation transfer functions for band-limited noise in subjects with cochlear hearing loss," Br. J. Audiol. **26**, 229–237.
Narayanan, A., and Wang, D. L. (**2012**). "A CASA-based system for long-term SNR estimation," IEEE Trans. Audio Speech Lang. Process. **20**, 2518–2527.
Nelson, P. B., and Thomas, S. D. (**1997**). "Gap detection as a function of stimulus loudness for listeners with and without hearing loss," J. Speech Lang. Hear. Res. **40**, 1387–1394.
Nilsson, M., Soli, S. D., and Sullivan, J. A. (**1994**). "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099.
Souza, P. E., and Boike, K. T. (**2006**). "Combining temporal-envelope cues across channels: Effects of age and hearing loss," J. Speech Lang. Hear. Res. **49**, 138–149.
ter Keurs, M., Festen, J. M., and Plomp, R. (**1992**). "Effect of spectral envelope smearing on speech reception I," J. Acoust. Soc. Am. **91**, 2872–2880.
ter Keurs, M., Festen, J. M., and Plomp, R. (**1993**). "Effect of spectral envelope smearing on speech reception II," J. Acoust. Soc. Am. **93**, 1547–1552.
Turner, C. W., Chi, S.-L., and Flock, S. (**1999**). "Limiting spectral resolution in speech for listeners with sensorineural hearing loss," J. Speech Lang. Hear. Res. **42**, 773–784.
Wang, D., and Brown, G. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms and Applications* (Wiley-IEEE, Hoboken, NJ), pp. 1–44.
Wang, D., Kjems, U., Pedersen, M., Boldt, J., and Tunner, T. (**2009**). "Speech intelligibility in background noise with ideal binary time-frequency masking," J. Acoust. Soc. Am. **125**, 2336–2347.
Wang, D. L. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer, Norwell, MA), pp. 181–197.
Wang, Y., Han, K., and Wang, D. (**2013**). "Exploring monaural features for classification-based speech segregation," IEEE Trans. Audio. Speech Lang. Process. **21**, 270–279.
Wang, Y., and Wang, D. (**2013**). "Towards scaling up classification-based speech separation," IEEE Trans. Audio. Speech Lang. Process. **21**, 1381–1390.
Wilson, R. H., and Carhart, R. (**1969**). "Influence of pulsed masking on the threshold for spondees," J. Acoust. Soc. Am. **46**, 998–1010.